AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

Beyond the Unit Root Question: Uncertainty and Inference

Clayton Webb University of Kansas
Suzanna Linn Penn State University
Matthew J. Lebo University of Western Ontario

Abstract: *A fundamental challenge facing applied time-series analysts is how to draw inferences about long-run relationships (LRR) when we are uncertain whether the data contain unit roots. Unit root tests are notoriously unreliable and often leave analysts uncertain, but popular extant methods hinge on correct classification. Webb, Linn, and Lebo (WLL; 2019) develop a framework for inference based on critical value bounds for hypothesis tests on the long-run multiplier (LRM) that eschews unit root tests and incorporates the uncertainty inherent in identifying the dynamic properties of the data into inferences about LRRs. We show how the WLL bounds procedure can be applied to any fully specified regression model to solve this fundamental challenge, extend the results of WLL by presenting a general set of critical value bounds to be used in applied work, and demonstrate the empirical relevance of the LRM bounds procedure in two applications.*

Verification Materials: The data and materials required to verify the computational reproducibility of the results, procedures, and analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/ZBRTJH.

Researchers who use time-series data are often interested in long-run, dynamic relationships between some (set of) independent variable(s), X, and some process, y. For example, do the levels of unemployment, inflation, and consumer sentiment affect the level of presidential approval over time (Clarke and Stewart 1994; De Boef and Kellstedt 2004)? Traditional approaches to analyzing these kinds of relationships have three steps: (1) use pretests to identify the orders of integration and univariate properties of each time series, (2) choose an appropriate model and hypothesis-testing framework based on these results, and (3) draw inferences from the model. The fundamental problem with these approaches is that the educated guesses made in the first step can be wrong and the uncertainty of these decisions is not reflected in the steps that follow. Practitioners ignore this hidden dimension of uncertainty and, as a result, can draw mistaken conclusions about the existence of long-run relationships (LRRs).

The tests available to classify time series as stationary, fractionally integrated, or unit root processes are notoriously unreliable. A number of factors complicate standard pretesting procedures: (1) The tests have low power. (2) The decisions about which tests are appropriate depend on the analyst's knowing something about the underlying dynamics of the data (e.g., is there a trend?). (3) The tests require analysts to make correct decisions about lag length, bandwidth, and/or lag truncation. (4) The knife-edged decisions of how to categorize time series can hinge on the chosen level of significance. As a result, pretests often produce inconclusive and conflicting results that leave the analyst with difficult decisions.

These classification decisions have important consequences. In particular, the validity of extant methods hinges on correct classification of the time series under analysis. An analyst typically uses a model to test a hypothesis about a particular kind of equilibrium. But to draw correct inferences about LRRs, the analyst needs to

Clayton Webb is Assistant Professor of Political Science, University of Kansas, 1541 Lilac Lane, Room 504, Lawrence, KS 66045 (webb767@ku.edu). Suzanna Linn is Professor of Political Science, Penn State University, 320 Pond Lab, University Park, PA 16802 (slinn@la.psu.edu). Matthew Lebo is Professor of Political Science, University of Western Ontario, Room 4154, Social Sciences Centre, London, Ontario N6A 3K7 Canada (matthew.lebo@uwo.ca).

We thank the Editors of *American Journal of Political Science* and *Political Analysis* for coordinating the simultaneous submission and review of this article and Webb, Linn, and Lebo (2019). We are also grateful to Dave Armstrong and Paul Johnson for computing help on the simulation designs and to John Freeman, Paul Kellstedt, Patrick Kraft, and Mark Pickup for helpful comments on drafts of this article. We thank the Center for Research Methods and Data Analysis and the College of Liberal Arts and Sciences at the University of Kansas for access to their high-performance computer cluster, on which many of the calculations reported here were conducted.

American Journal of Political Science, Vol. 64, No. 2, April 2020, Pp. 275–292

 DOI: 10.1111/ajps.12506

know whether the data are stationary, contain unit roots, or are fractionally integrated.[1] Thus, the uncertainty surrounding pretests creates a fundamental challenge for time-series analysts: How should we analyze long-run relationships when we are uncertain whether the data are stationary or contain unit roots?

A recent flurry of articles has grappled with this problem. Grant and Lebo (2016), Lebo and Grant (2016), Keele, Linn, and Webb (2016a, 2016b), Esarey (2016), and Helgason (2016) demonstrate when standard strategies are likely to be vulnerable to misclassification and when they may be fairly robust. Enns et al. (2016), Lebo and Kraft (2017), and Enns et al. (2017) contend with the different substantive conclusions that come from divergent interpretations of unit root test results. Enns and Wlezien (2017) and Pickup and Kellstedt (2020) discuss the complications that arise when analysts conclude their regressors and regressands have different orders of integration. Yet, none of this work provides a general solution to the problem. Philips (2018) recommends adopting the cointegration testing strategy developed by Pesaran, Shin, and Smith (PSS; 2001) that allows the analyst to be agnostic about the properties of their regressors. While this procedure represents an important innovation, the approach developed by PSS requires that the analyst know the univariate properties of the regressand and breaks down if y is not a unit root (Webb, Linn, and Lebo 2019).

What, then, should analysts do if they are uncertain whether their data are stationary or contain unit roots? Webb, Linn, and Lebo (WLL; 2019) present an extensive analytical critique of the PSS procedure and develop an alternative framework for inference that resolves the limitations of the PSS approach. It uses critical value bounds for hypothesis tests on the long-run multiplier to incorporate the uncertainty inherent in identifying the dynamic properties of *all* the data into inferences about LRRs. The testing strategy is straightforward and highly generalizable—it can be applied to any fully specified dynamic regression model.

In this article, we demonstrate how this framework presents a solution to the fundamental challenge of time series. In the next section, we describe the two central problems complicating tests for long-run relationships: the practical problems of pretesting and the ambiguities in extant tests for LRRs given this uncertainty. We then discuss concepts of equilibrium and show the problem with tests commonly applied in the GECM framework. Following that, we present both the long-run multiplier (LRM) t-test as a general test for long-run relationships and a bounds framework for assessing the significance of the LRM test statistic that accommodates uncertainty about the univariate properties of the data. We extend the results of WLL to cover a wider range of conditions likely to be relevant in applied time series analysis. Finally, we provide examples of the procedure for both the autoregressive distributed lag and generalized error correction models to demonstrate the generalizability of the test and the empirical relevance of the procedure.

The Practical Problems of Pretesting

Testing for unit roots is complicated. A variety of tests have been proposed to help the analyst determine whether a time series contains a unit root. Most are based on an autoregressive representation of the series:

$$y_t = D_t + \rho y_{t-1} + \mu_t,$$

where D_t captures the deterministic features of the process, generally a constant or linear trend, and μ_t is a white noise process.[2] The hypothesis of interest concerns ρ. If $\rho = 1$, y contains a unit root. Most tests evaluate the unit root null hypothesis. The (Augmented) Dickey-Fuller test (1979), the Phillips-Perron (1988) family of tests, and the DF-GLS test (Elliott, Rothenberg, and Stock 1996) all test the unit root null hypothesis. Alternatively, the KPSS test (Kwiatkowski et al. 1992) tests the null hypothesis of stationarity.[3] All these tests have nonstandard limiting distributions, which depend on the form of D_t—whether a constant, a trend, or neither is included in the test regression.

There is broad agreement in the econometrics literature that unit root tests have low power (Banerjee et al. 1993; Campbell and Perron 1991; DeJong et al. 1992; Elliott, Rothenberg, and Stock 1996; Evans and Savin 1981, 1984; Juhl and Xiao 2003; Perron and Ng 1996;

[1] This is true whether we estimate the popular autoregressive distributed lag model (ADL), use the generalized error correction model (GECM; De Boef and Granato 1997; De Boef and Keele 2008; Engle and Granger 1987; Ericsson and MacKinnon 2002), or model the series as a fractionally integrated process (Box-Steffensmeier and Tomlinson 2000; Clarke and Lebo 2003; Lebo and Grant 2016).

[2] A general model for a single time series can be defined as follows: $(1 - \sum_{k=1}^{p} \rho_k L^k)(1 - L)^d y_t = (1 + \sum_{k=1}^{q} \theta_k L^k)\epsilon_t + \tau$. The current value y_t is a function of p autoregressive parameters (ρ_p), q is moving average parameters (θ_q), ϵ_t is a white noise error term, L is the lag operator such that $L^k y_t = y_{t-k}$, τ is a trend, and the differencing parameter, d, tells us how many times the series must be differenced to make it stationary. A weakly stationary series exhibits mean reversion ($\mu = E(y_t) = E(y_{t+s})$), finite variance ($\sigma_y^2 = E[(y_t - \mu)^2] = E[(y_{t+s} - \mu)^2]$), and stationary covariance ($\gamma = E[(y_t - \mu)(y_{t+s} - \mu)] = E[(y_{t-j} - \mu)(y_{t-j+s} - \mu)]$ for all s).

[3] See Enders (2015) or Box-Steffensmeier et al. (2014) for a thorough discussion of unit root tests.

Stock 1991).[4] Simply put, the problem is that "it is difficult for *any* statistical procedure to distinguish between unit root processes and series that are highly persistent" (Enders 2015, 235; emphasis added).[5] Many time series of interest to political scientists, particularly those measuring public opinion, economic conditions, and budgeting, tend to be characterized by inertia. Series often appear to behave like unit roots. This makes the low power of unit root tests problematic, particularly in small samples.

Unit root tests also rely on the correct specification of D_t. "Inappropriately omitting the intercept or time trend can cause the power of the test to go to zero.... On the other hand, extra regressors increase the critical values so that you may fail to reject the null of a unit root" (Enders 2015, 235).[6] Omitting a constant from a test imposes the assumption that the mean of the series is zero; omitting a trend imposes the assumption that the mean of the series is constant over time. In many cases, the existence of a deterministic trend may seem implausible, but the inclusion or omission of a trend term can have an outsized influence on the results of different testing procedures. If the analyst is agnostic, auxiliary tests can help select the appropriate form of D_t. However, these procedures have their own problems. Enders (2015, 237) emphasizes that "tests for unit roots are conditional on the presence of deterministic regressors and tests for the presence of deterministic regressors are conditional on the presence of a unit root." Theory and visual inspection of a time series can aid in decision making, but the appropriate form of D_t is not always clear.

Still another complication is presented by serially correlated errors. The Dickey-Fuller test assumes the residuals are white noise. The analyst can *augment* the test to accomodate serial correlation by adding lags of Δy_t, but the analyst must select the correct lag length. If the analyst includes too many lags, the power of the test suffers. If the analyst includes too few, the size of the test suffers. One might use information criteria to select the best lag length, but Ng and Perron (2001) show that standard criteria do not maximize test power. The KPSS and Phillips-Perron tests can accommodate serial correlation via nonparametric estimates of the long-run variances, but the power and size of these tests depend on the bandwidth and lag truncation parameters used in the calculation of the variances. Unfortunately, there are no universally optimal strategies for selecting these elements of the tests.

There are a host of other considerations that affect unit root tests. Assumptions about the initial observation of a series and about the presence, or absence, of structural breaks also affect inferences (e.g., see Müller and Elliott 2003; Perron 1989). The level of aggregation and sampling window are important as well. Smaller samples, like those often available in political science, make it less likely to correctly reject the unit root null hypothesis. Popular software packages handle the testing process differently, with varying automated methods for choosing lag lengths and different algorithms for interpolating critical values. Finally, there are also theoretical considerations that may weigh on classification. For example, Williams (1992) argues that series with upper and lower limits cannot have infinite variance and, therefore, cannot be unit roots.

Given the complications of unit root testing, the recommended strategy is to conduct multiple tests, including a series of auxiliary tests designed to identify the correct form of D_t. But existing empirical strategies often produce inconsistent results and lead analysts to select the evidence that best suits their preferred conclusions. Inconsistency among tests generates a profound level of uncertainty in applied research that is not reflected in final analyses.

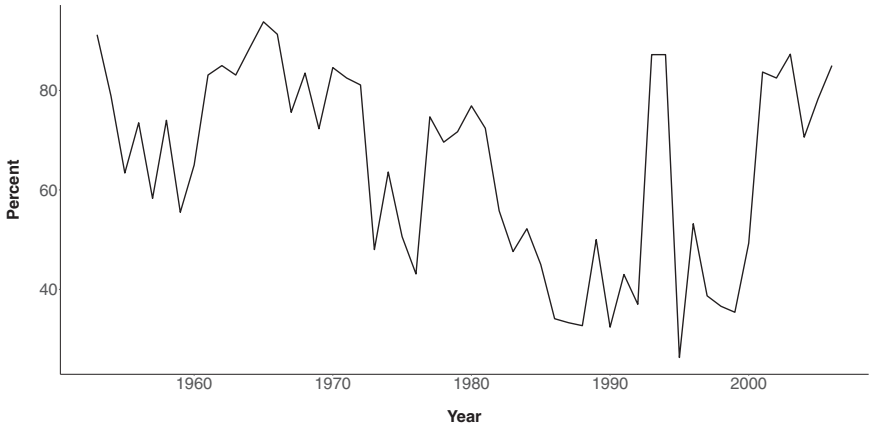
To illustrate this problem, we examine *Presidential Success*, the yearly percentage of votes on bills on which the president has taken a winning position in the House of Representatives from 1953 to 2006 (Lebo and O'Geen 2011). The series is typical of studies involving American institutions—it has a small sample size and is restricted between upper and lower limits. A theoretical argument could be made that the parties of the president and the House majority fall in and out of sync over time, and this should cause the series to be mean reverting. But with presidents facing the same Congress for 2 years at a time, enough autocorrelation may exist in the yearly data to make it difficult to reject the unit root null hypothesis.

[4]Even more problematic, "no single test is uniformly most powerful" (Choi 2015, 52), and those with greater power tend to have relatively poor size. Relative power varies across different classes of data generating processes (DGPs) and given different assumptions about the first observation. Even optimal tests have arbitrarily low power against local alternatives, and stationarity tests have parallel size problems (Choi 2015, 127).

[5]Even when the analyst has a large sample, D_t is known, there are no structural breaks, the DGP has moderate autocorrelation, and the errors are well behaved, Elliott, Rothenberg, and Stock (1996) show that power of the optimal test is poor. As the sample size shrinks and the DGP becomes more strongly autoregressive, power drops substantially. In a sample of size 50 with $\rho = .85$, for example, the optimal unit root test will correctly reject the null hypothesis less than 20% of the time (Podivinsky and King 2000).

[6]The appropriate form of D_t is that specified under the *alternative* hypothesis. For example, in a Dickey-Fuller regression that includes a constant and trend, under the null hypothesis the trend term is assumed to be zero; that is, the series is a unit root with drift. Under the alternative, the trend is significant and the process is trend stationary. Similarly, if D_t includes a constant only, it is assumed to be zero under the null hypothesis such that the process is a pure random walk under the null. Under the alternative hypothesis, the series is mean stationary (Choi 2015, 30).

FIGURE 1 Presidential Success in the House of Representatives, 1953–2006



Note: *Presidential Success* is the yearly percentage of votes on bills on which the president has taken a winning position in the House of Representatives from 1953 to 2006 (Lebo and O'Geen 2011).

The series is presented in Figure 1. It appears stationary, moving around a long-run mean of 68.8%.

We conduct a battery of commonly used tests to infer whether presidential success is a unit root or stationary process and present the results in Table 1.[7] The tests are organized by D_t, the null and alternative hypotheses tested. We present results for several lag lengths, including the "short" and "long" lag truncations recommended by Schwert (1989) ($trunc(4 * (n/100)^{0.25}) = 3$ and $trunc(12 * (n/100)^{0.25}) = 9$), the lag lengths selected by the general-to-specific modeling strategy for the Dickey-Fuller and DF-GLS tests, and those selected using the Akaike information criterion (AIC) for the Dickey-Fuller tests. We also present the results for the KPSS test with no lags and results for the Phillips-Perron tests with one lag. The number of lags used in each case is given in parentheses as appropriate.

This empirical strategy yields inconsistent results. For test regressions with (1) a constant and trend or (2) a constant only, the Dickey-Fuller tests reject the unit root null if one adopts a general-to-specific strategy for lag selection but fail to do so if the AIC is used.[8] If one adopts the short lag, the Dickey-Fuller and DF-GLS tests fail to reject the unit root null at the $\alpha = .05$ level (but the DF-GLS rejects it at the .10 level when D_t includes only a constant), but inferences from both versions of the Phillips-Perron tests would lead the analyst to infer the series is stationary (around a trend or constant). Lag selection is also pivotal in determining the inferences from the KPSS tests. If one uses the "nil" option, both tests present strong evidence against the stationary null, but using either three or nine lags, one cannot reject the null at the $\alpha = .05$ level. Adopting $\alpha = .10$, however, one would reject the stationary null, with the exception of the null that the series is mean stationary at lag 9. In short, even when following recommended best practices for unit root testing, inferences are highly uncertain.

The difficulties of pretesting are not unique to presidential success but abound in political science. Popular time series such as approval ratings, public policy mood, consumer sentiment, and foreign policy conflict have been treated as stationary by some analysts and as unit root processes by others. As we discuss next, this uncertainty presents problems for drawing inferences about LRRs.

Ambiguities in Testing for LRRs

All single-equation dynamic regression models specify an LRR between X and y.[9] These LRRs imply the existence of long-run equilibria. An equilibrium state exists when

[7] We implemented the tests using the functions provided in the `urca` package in `R`.

[8] We fail to reject the null that the data are generated by the restricted form of the test regression with no trend and no constant (Dickey-Fuller ϕ_2) and with no trend (Dickey-Fuller ϕ_3) for all lags using an $\alpha = .05$. However, in a test with no lags, we can reject both ϕ_2 and ϕ_3 at $\alpha = .10$. Using ϕ_1, we reject the null hypothesis that a constant can be dropped from the test regression in versions of the test with no lags, but we fail to do so for all other lag lengths.

[9] We use X to denote a set of regressors and x to denote a single regressor.

TABLE 1 Unit Root and Stationarity Tests: Presidential Success in the U.S. House, 1953–2006

	Test Statistic				
Test[a]	**General to Specific**	**AIC**	**Short Lag 3**	**Long Lag 9**	**Alternative**
$D_t = (1, t)$; Null: unit root with drift; Alternative: trend stationary					
Dickey-Fuller τ_τ	$-3.51^{*}(0)$	$-2.29(1)$	-2.22	-0.94	
DF-GLS	$-3.49^{*}(0)$		-2.33	-1.37	
Phillips-Perron Z_τ			-3.56^{*}	-3.63^{*}	$-3.44^{+}(1)$
Phillips-Perron Z_α			-22.83^{*}	-24.00^{*}	$-21.21^{*}(1)$
$D_t = (1, 0)$; Null: unit root; Alternative: mean stationary					
Dickey-Fuller τ_μ	$-3.53^{**}(0)$	$-2.45(1)$	-2.23	-1.37	
DF-GLS	$-2.77^{**}(0)$		-1.69^{+}	-0.85	
Phillips-Perron Z_τ			-3.54^{*}	-3.72^{**}	$-3.45^{*}(1)$
Phillips-Perron Z_α			-20.55^{**}	-23.24^{**}	$-19.25^{**}(1)$
$D_t = (0, 0)$; Null: unit root; Alternative: mean-zero stationary					
Dickey-Fuller τ	$-0.09(6)$	$-0.67(1)$	-0.39	-0.32	
$D_t = (1, t)$; Null: trend stationary; Alternative: not trend stationary					
KPSS τ			0.14^{+}	0.12^{+}	$0.33^{**}(0)$
$D_t = (1, 0)$; Null: mean stationary; Alternative: not mean stationary					
KPSS μ			0.44^{+}	0.27	$1.11^{**}(0)$

Note: $T = 54$. All tests were conducted in R using the urca package. Short lag truncation is based on the formula $trunc(4 * (n/100)^{0.25}) = 3$; long lag truncation is based on the formula $trunc(12 * (n/100)^{0.25}) = 9$ as given in Schwert (1989) and used in urca. As suggested by Ng and Perron (1995), the longer lag length was used as the maximum lag length for selecting the appropriate lag length for the Dickey-Fuller and DF-GLS tests. The number of lags chosen based on a general-to-specific modeling strategy ($\alpha = .05$ was used as the cut-off) and using the AIC is given in parentheses. The column labeled "Alternative" includes results for (a) Philips-Perron tests using one lag, a common specification choice, and (b) the KPSS test assuming no lag lengths, as this is an option in urca.

[a]The form of D_t for unit root tests is specified under the alternative hypothesis.

$^{**}p < .01$, $^{*}p < .05$, $^{+}p < .10$.

the variables in a system exhibit no tendency to change over time (Banerjee et al. 1993, 2). We do not observe social phenomena in their equilibrium states because social phenomena are rarely at rest. Instead, we observe equilibria as levels or relationships to which variables tend to return. Series may deviate from their equilibria, but they are unlikely to do so by very much or for very long (Burke and Hunter 2005, 38).

Equilibria can take various forms. An individual time series may or may not have an equilibrium. A unit root series, unrelated to any other measured variables, wanders unpredictably. Without reference to another time series, it cannot be said to be in either equilibrium or disequilibrium at any given point in time. A unit root series may have an equilibrium in reference to another unit root variable or variables. If both y and X contain unit roots but their linear combination creates a mean-reverting series, there is a long-run *cointegrating equilibrium*. Equilibrium implies that y and at least one element of X are aligned, and disequilibrium indicates y's separation from at least one element of X.

A stationary time series always has an equilibrium; it returns to a long-run mean whether it is related to a given set of exogenous regressors or not. If a stationary y is not related to X, its equilibrium is determined by variables not included in the model. We call this an *unconditional stationary equilibrium*. In contrast, a *conditional stationary equilibrium* exists when y is a function of X. In each case, disequilibrium implies y is away from its long-run mean and mean-reverting behavior is soon to follow. In many applications, however, it is ambiguous which type of equilibrium is at work when researchers have claimed to find LRRs.

In "Taking Time Seriously" (TTS), De Boef and Keele (2008) point out that since the ADL and GECM are mathematically equivalent, the GECM can be used for either nonstationary or stationary data. Thus, a common mistake is to use TTS as a license to ignore the univariate

properties of one's data. Grant and Lebo (2016) demonstrate problems from hasty interpretation of the GECM. Keele, Linn, and Webb (2016b) attempt to clarify the message in TTS, explaining that the interpretation of the GECM's error correction coefficient, α_1^*, depends on analysts' conclusions about whether their series are individually integrated. Despite this recent exchange, however, confusion remains about how to conduct hypothesis tests for LRRs.

Indeed, extant tests for long-run equilibria require analysts to know the univariate properties of their data, since the null and alternative hypotheses and the appropriate critical values depend on the properties of y. But without confidence in the results from these pretests, analysts cannot know which types of equilibria are possible, which models to estimate, which tests should be applied, or how to interpret test results. If analysts misclassify y, they will reach incorrect inferences about equilibria. This is the fundamental problem with traditional approaches. Analyses depend on uncertain first steps, and this uncertainty is not reflected in published results.

We can illustrate the problem with a basic dynamic model. We begin with the error correction model (ECM) because it isolates the LRR explicitly (Hendry 1995; Pagan 1987). Consider the simple bivariate case:

$$\Delta y_t = \alpha_0 + \beta_0^* \Delta x_t + \alpha_1^*(y_{t-1} - \lambda x_{t-1}) + e_t, \quad (1)$$

where λ is the long-run multiplier, giving the total effect of x on y distributed over time; $y_{t-1} - \lambda x_{t-1}$ identifies the long-run equilibrium relationship and measures the disequilibrium between x and y; and α_1^*, the familiar error correction coefficient, gives the rate of return to equilibrium after a shock.

This model is often estimated as a *generalized* error correction model (GECM), which distributes terms:

$$\Delta y_t = \alpha_0 + \beta_0^* \Delta x_t + \alpha_1^* y_{t-1} + \beta_1^* x_{t-1} + e_t, \quad (2)$$

or as an autoregressive distributed lag model (ADL):

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + e_t. \quad (3)$$

All three representations are equivalent. Specifically, the impact multipliers in the ECM are equivalent to those in the GECM and ADL: $\beta_0^* = \beta_0$ and $\alpha_1^*\lambda = \beta_1^* = \beta_0 + \beta_1$; the LRM, λ, in the ECM is equal to $-\frac{\beta_1^*}{\alpha_1^*}$ in the GECM and to $\frac{\beta_0+\beta_1}{1-\alpha_1}$ in the ADL; and the constant, α_0, is given by the same coefficient in each representation.

Consider Equation (2). If the analyst is certain the data are all unit roots, inference about the existence of a cointegrating LRR may be drawn from a hypothesis test that $\alpha_1^* = 0$, which tests the null hypothesis that the two series are not cointegrated against the alternative that they are cointegrated. The t-statistic is nonstandard in this case, and appropriate critical values are given in Ericsson and MacKinnon (2002). Finding that $\alpha_1^* = 0$ in Equation (2) implies that λ in Equation (1) is undefined.[10] In contrast, a nonzero α_1^* implies that λ and β_1^* are nonzero. This is true because if x did not define the long-run equilibrium in y, no such equilibrium would exist. Thus, rejection of the null $\alpha_1^* = 0$ is sufficient for inferring a cointegrating relationship between x and y when the data are certain to contain unit roots.[11]

In the stationary case, testing the null hypothesis on α_1^* relies on an asymptotically normal distribution and is no longer a test for cointegration (Banerjee et al. 1993, 167). The null hypothesis $H_0 : \alpha_1^* = 0$ is that y has no long-run equilibrium—in which case y is a unit root process—and the alternative hypothesis is that y has an equilibrium. The alternative is trivially true for autoregressive processes. Given a stationary y, α_1^* must be nonzero *irrespective of y's relationship with X*, in the same way α_1 can be nonzero in the ADL and yet unrelated to X. Put differently, α_1^* in the GECM will be nonzero regardless of whether λ and β_1^* are also nonzero, that is, regardless of whether y has an unconditional stationary equilibrium or an equilibrium value conditional on X.

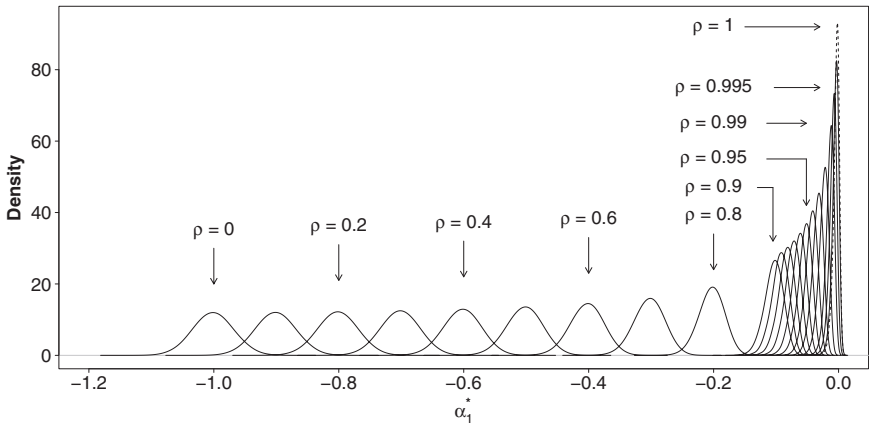
Figure 2 shows the distribution of α_1^* simulated from a GECM where y has varying levels of autocorrelation and is not related to an x. Clearly, the value of α_1^* is linearly related to the level of autocorrelation in y. When y is a simple white noise process ($\rho = 0$), the values of α_1^* center on -1. With $\rho = 0.5$, the values of α_1^* center on -0.5. This makes sense because α_1^* is capturing the speed of mean reversion. Relating this to Equation (1), if data are stationary, a nonzero and significant α_1^* can exist when both $\lambda = 0$ and $\beta_1^* = 0$. That is, analysts will almost always reject the null hypothesis with respect to the "error correction coefficient" whether or not y is a function of X. Thus, a crucial problem is that inference based on α_1^* in the GECM (or α_1 in the ADL) cannot discern between conditional and unconditional equilibration absent certainty about the univariate properties of the data. Confusion over this distinction has been evident in many applications of the GECM.

Analysts can only know how to interpret the hypothesis tests on α_1^* when they know the properties of the data with certainty. With uncertainty about the nature of our data, tests based on $\alpha_1^* = 0$ in the GECM do not

[10] If α_1^* is exactly 0, λ in undefined. In applied settings, α_1^* will not be exactly zero, but the analyst will fail to reject the null that $\alpha_1^* = 0$ and the null that $\lambda = 0$.

[11] In the case with two or more regressors, λ and β_1^* are vectors, and a significant α_1^* implies a cointegrating relationship between y and *at least one* regressor, such that some element of λ and β_1^* are nonzero.

FIGURE 2 Sampling Distribution of α_1^* by Degree of Autocorrelation



Note: The distributions were produced from 50,000 simulations of the GECM for $y_t = \rho_y y_{t-1} + \varepsilon_t$ and $\rho_y = 0.0$ to 0.9 in steps of 0.1; $\rho_y = 0.91$ to 0.99 in steps of 0.01; $\rho_y = 0.995$; and $\rho_y = 0.999$.

conclusively imply an LRR between y and any element of X. Rejecting the null hypothesis $\alpha_1^* = 0$ might imply cointegration or it might not. A significant α_1^* might measure the rate of error correction *between* y and X or it might not; it might simply indicate the stationarity of y. Lastly, finding a nonzero α_1^* might imply that λ and β_1^* must also be nonzero or it might not. The correct interpretation of α_1^* depends on the univariate properties of the series (Banerjee et al. 1993, 167).

As Webb, Linn, and Lebo (2019) demonstrate, this ambiguity also affects the interpretation of alternative tests for LRRs, including the widely cited tests proposed by Pesaran, Shin, and Smith (2001). The PSS testing procedure applies critical value bounds to the conventional t-test on α_1^*, as well as an F-test for cointegration, to accommodate pretest uncertainty in X.[12] PSS's bounds encompass all possible critical values that could be correct given any possible combination of unit root and stationary independent variables. The analyst compares the test statistic to the upper and lower limits. If the test statistic is beyond the upper bound, the analyst can confidently reject the null hypothesis, regardless of the univariate properties of X. The analyst's uncertainty about the univariate dynamics in X is reflected in the area of indeterminacy between the bounds. As Philips (2018, 230) stated when introducing the method to political scientists, the "strategy absolves users from having to distinguish between stationary . . . and first-order non-stationary ($I(1)$) regressors. This is an advantage since unit-root testing is difficult in short series, and introduces 'a further degree of uncertainty into the analysis' (PSS, p. 289)." However, as we discussed earlier, it may be difficult to confidently conclude y is $I(1)$. Without the ironclad assurance that y is $I(1)$, the entire PSS hypothesis-testing procedure breaks down in the same fashion illustrated above: The test statistics will be nonzero if y is stationary, regardless of whether there is an LRR between X and y. We need a framework for testing for LRRs that does not rely on analysts knowing the univariate properties of any of the time series in their models.[13]

A General Test for Long-Run Relationships and a Bounds Approach to Inference

Webb, Linn, and Lebo (2019) offer a solution to the problem that builds on the logic of the PSS approach.

[12]Of course, it is inappropriate to test for cointegration if y is not a unit root. All cointegration tests break down if the analyst is not certain that y is $I(1)$.

[13]Bayesian approaches present an alternative that allows beliefs about the existence of unit roots (or near and fractional integration) and cointegration to be "expressed as probabilistic statements rather than based on knife-edge tests" (Brandt and Freeman 2009, 124). In a structural Bayesian vector autoregression, these beliefs are elucidated based on theory, translated into hyperparameters for estimation, and then assessed via sensitivity analysis and for the consistency of forecasts with prior knowledge. As Brandt and Freeman (2009, 124) note, this approach means the "analyst need not perform any pre-tests that could produce mistaken inferences about the trend properties of" their data.

Specifically, they establish critical value bounds to reflect pretest uncertainty in tests for LRRs based on the long-run multiplier. The authors show that the LRM estimated from any correctly specified dynamic model will only be defined and nonzero if there is a conditional stationary or conditional cointegrating LRR between a given x and y. They demonstrate the distributional properties of the LRM t-statistic and use simulations to derive the bounds for the t-test on the LRM. We explain the intuition of their statistical findings and discuss the strengths and weaknesses of adopting their approach to testing for LRRs. For the reader interested in the technical details, see Webb, Linn, and Lebo (2019).

Webb, Linn, and Lebo (2019) propose using the LRM to test for the existence of LRRs. The logic can be easily understood by considering the LRM in the simple bivariate GECM presented in Equation (2), where $\lambda = -\frac{\beta_1^*}{\alpha_1^*}$. The LRM gives the total effect of a unit change in x on y. A conditional LRR therefore will only exist if λ is defined and nonzero. This requires both the numerator and denominator to be nonzero. Consider the case where both x and y are unit roots. Cointegration requires that $\alpha_1^* \neq 0$, which can only occur if y is linked to x (i.e., the numerator is also nonzero). If there is no long-run cointegrating relationship, $\alpha_1^* = 0$, there is no link to x in the long run, and λ is undefined. If the data are stationary, λ is always defined because α_1^* is by definition nonzero. But λ will only be nonzero if y is linked to x, which implies the numerator is also nonzero and there is a long-run stationary relationship. If y is not related to x in the long run, $\lambda = 0$. The logic extends to the model where y is a function of X and holds for any dynamically complete model specification, including the ADL (Equation 3).

Thus, the test of the significance of each LRM in any dynamic regression is a test of a conditional LRR between x and y, regardless of the univariate properties of the data. Failure to reject the null means one cannot conclude an LRR exists, regardless of whether the data are unit root or stationary processes. Rejecting the null means we can infer a long-run equilibrium relationship between x and y. *This is the test we want.*

The LRM and its standard error are not directly estimated in the GECM (or in the ADL). One can, however, calculate the LRM from the estimates in any dynamic regression using the appropriate formula for the LRM. For the ADL, $\lambda = \frac{\beta_0+\beta_1}{1-\alpha_1}$. For the GECM, $\lambda = -\frac{\beta_1^*}{\alpha_1^*}$. The resulting estimate gives us the total effect of an x on y in the long run. The standard error of the LRM can be calculated using the delta method or by estimating the LRM and its standard error directly using the Bewley transformation of the original model (Bewley 1979):

$$y_t = \phi_0 - \phi_1 \Delta y_t + \psi_0 X_t - \psi_1 \Delta X_t, + \mu_t, \qquad (4)$$

where ψ_1 is the LRM.[14]

De Boef and Keele (2008) recommend that practitioners use the LRM to draw inferences about LRRs but do not consider how the dynamic properties of the data affect the appropriate critical values for hypothesis tests. However, as Webb, Linn, and Lebo (2019) demonstrate, the distribution of the LRM t-statistic, and thus the critical values for the hypothesis test on the LRM, depends on the univariate characteristics of the variables in the model, as well as the deterministic features of the DGP for y, the number of independent variables, and the sample size. The authors use this information to establish critical value bounds for the LRM t-test that reflect pretesting uncertainty. These are the highest and lowest critical values associated with a given confidence level for the LRM t-test given (1) any possible degree of autocorrelation in either y or X, between and including $\rho = 0$ and $\rho = 1$;[15] (2) the presence or absence of cointegrated X; (3) the presence or absence of a trend in the DGP of y; and (4) the presence or absence of a constant in the DGP for y.[16] As such, the critical value bounds they identify accommodate

[14]The Bewley coefficients are linear transformations of the GECM and ADL coefficients. In Equation (2), $\phi_0 = -\frac{\alpha_0}{\alpha_1^*}$, $\phi_1 = -\frac{\alpha_1^*+1}{\alpha_1^*}$, $\psi_0 = \beta_1^*$, $\psi_1 = -\frac{\beta_1^*}{\alpha_1^*}$, and $\mu = -\frac{e}{\alpha_1^*}$. Translating from the ADL in Equation (3), $\phi_0 = \eta\,\alpha_0$, $\phi_1 = \eta\,\alpha_1$, $\psi_0 = \eta(\beta_0 + \beta_1)$, $\psi_1 = \eta\,\beta_1$, $\mu = \eta\,\varepsilon_t$, and $\eta = \frac{1}{\alpha_1 - 1}$. A constant, X_t, X_{t-1}, and y_{t-1} should be used as instruments to estimate the model (De Boef and Keele 2008). If a trend is part of the DGP, then a trend term should be included in the Bewley transformation as well as the GECM or ADL.

[15]We can set aside concerns about fractional integration. The bounds encompass $d = 0$ to $d = 1$.

[16]Details are given in Webb, Linn, and Lebo (2019). Briefly, critical values are computed via stochastic simulations under the true null hypothesis of no LRR between X and y. A number of experiments were conducted in which y is generated as an autoregressive process ($y_t = c_0 + c_1 t + \rho_y y_{t-1} + e_{y,t}$) independent of X. The degree of autocorrelation and the presence or absence of a constant and trend are varied across the experiments. X is generated under a range of conditions, including (1) the case where all the elements of X are independent $I(0)$ processes ($x_{k,t} = \rho_{x_k} x_{k,t-1} + e_{x_k,t}, 0 \leq \rho < 1.0$), (2) all are independent unit root processes ($x_{k,t} = x_{k,t-1} + e_{x_k,t}$), and (3) any number of x are cointegrated ($x_{1,t} = x_{1,t-1} + e_{x_1,t}$, and $x_{k,t} = \rho_{x_k} x_{k,t-1} + e_{x_k,t}$ for some $k \neq 1$). For each experimental condition, the LRM, its standard error, and the associated t-statistic were estimated from the Bewley model. The experiment was repeated 100,000 times. The quantiles of the resulting distributions provide critical values for each characterization of the DGPs examined. These are used to set the bounds by identifying the smallest and largest critical values that are estimated across the experiments.

TABLE 2 Upper Bounds (UB) and Lower Bounds (LB) for the LRM *t*-Test by *k*, *T*, and α

	$T = 25$		$T = 50$		$T = 75$		$T = 150$		$T = 500$		$T = 1,000$	
k	**LB**	**UB**	**LB**	**UB**	**LB**	**UB**	**LB**	**UB**	**LB**	**UB**	**LB**	**UB**
$\alpha = .01$												
1	1.84	6.28	1.54	6.16	1.46	5.98	1.36	5.94	1.32	5.87	1.31	5.88
2	1.91	6.19	1.53	6.06	1.45	5.90	1.37	5.93	1.32	5.85	1.31	5.88
3	2.00	6.15	1.56	5.96	1.45	5.83	1.37	5.89	1.32	5.82	1.31	5.87
4	2.10	6.10	1.57	5.97	1.45	5.77	1.38	5.89	1.32	5.83	1.31	5.83
5	2.25	6.05	1.57	5.85	1.47	5.74	1.37	5.86	1.32	5.81	1.31	5.84
$\alpha = .05$												
1	1.25	3.79	1.09	3.72	1.06	3.73	1.01	3.68	0.99	3.66	0.99	3.62
2	1.27	3.72	1.10	3.68	1.06	3.70	1.02	3.67	0.99	3.66	0.99	3.62
3	1.29	3.68	1.10	3.64	1.06	3.65	1.01	3.65	0.99	3.65	0.99	3.62
4	1.33	3.57	1.11	3.57	1.06	3.62	1.01	3.63	0.99	3.64	0.99	3.61
5	1.39	3.50	1.11	3.52	1.07	3.59	1.01	3.61	0.99	3.64	0.99	3.62
$\alpha = .10$												
1	0.99	2.84	0.89	2.83	0.87	2.81	0.84	2.79	0.83	2.76	0.83	2.78
2	1.00	2.76	0.89	2.77	0.87	2.79	0.85	2.77	0.83	2.77	0.83	2.78
3	1.02	2.70	0.90	2.73	0.87	2.77	0.84	2.76	0.83	2.76	0.83	2.78
4	1.05	2.64	0.90	2.68	0.88	2.73	0.84	2.75	0.83	2.75	0.83	2.77
5	1.08	2.55	0.91	2.64	0.88	2.70	0.84	2.73	0.83	2.75	0.83	2.77

Note: The 90%, 95%, and 99% critical values are computed via stochastic simulations using 100,000 replications for the LRM *t*-statistics in Equation (4), the Bewley IV regression. The time series y and X are generated from $y_t = \rho_y y_{t-1} + e_{yt}$ and $x_{i,t} = \rho_{x_i} x_{i,t-1} + e_{x_i,t}$ for $k = 1, 2, 3, 4, 5$ regressors where the errors are drawn from independent standard normal distributions.

> every type of analytical uncertainty that typically vex time series analysts. One does not need to know whether a series is a stationary or unit root process. One does not need to know whether a series is characterized as a random walk, a random walk with drift, or a random walk with trend and drift. These bounds allow analysts to focus on the theoretical questions at the heart of political analysis, the existence of LRRs. (Webb, Linn, and Lebo 2019, 14)

We extend the bounds calculated by Webb, Linn, and Lebo (2019) for the LRM *t*-statistic to cover six sample sizes, $T = \{25, 50, 75, 150, 500, \text{and } 1{,}000\}$, for models containing $k = 1$ to 5 independent variables, and for $\alpha = \{0.01, 0.05, 0.10\}$. The results are presented in Table 2. Note that the bounds are highly stable across sample sizes and the number of independent variables, with the important exception of very small samples.

The testing procedure is generalizable and straightforward. The analyst estimates an appropriately general dynamic regression model and calculates the LRM and its standard error using either the delta method or the Bewley transformation. The analyst computes the LRM test statistic and compares the absolute value to the appropriate bounds given the length of the series (T) and the number of independent variables (k). If the test statistic is below the lower bound, the analyst fails to reject the null hypothesis of no LRR. If the test statistic is above the upper bound, the analyst can reject the null hypothesis of no LRR with confidence. If the test statistic is between the bounds, the analyst cannot draw a firm conclusion. The area of indeterminacy between the bounds reflects the analyst's uncertainty about the true nature of *all* the series in the analysis.

Incorporating uncertainty into a hypothesis-testing procedure has consequences. First, the area of indeterminacy means that analysts may find themselves with inconclusive evidence about the existence of an LRR. Analysts are used to applying knife-edge logic to hypothesis tests. Applying a procedure where a possible outcome is "I can't be sure" is unusual and may seem unsatisfying. Second, although analysts can have confidence that an LRR exists if the LRM test statistic is beyond the upper bound, they cannot know *which type* of LRR exists: The test cannot distinguish an LRR between two unit root processes—cointegration—from an LRR between two stationary variables, a conditional stationary

equilibrium. Analysts need to have knowledge about the univariate properties of their series to differentiate stationary conditional equilibria from cointegrating equilibria. Finally, there is a greater chance for type II errors in some circumstances. If the analyst is *certain* the data are stationary but chooses to apply the bounds procedure and finds a value of the test statistic between the standard critical value and the upper bound, the analyst will incorrectly conclude that the dependent variable is defined by an unconditional equilibrium instead of a conditional stationary equilibrium.

If the analyst has accepted *ex ante* an inability to make reliable decisions about the properties of his or her data, these consequences are benefits, not costs. The area of indeterminacy may seem strange, but the area between the bounds reflects the uncertainty that is an inherent part of pretesting. If the analyst is unable to make reliable decisions about stationarity, this is an acceptable, and necessary, sacrifice. Importantly, the uncertainty inherent in the area of indeterminancy and about the type of LRR has always been a hidden part of alternative procedures. Rather than glossing over that uncertainty, the bounds procedure provides a principled basis for inference that makes the consequences of pretest uncertainty explicit. Moreover, there is nothing to prevent analysts from applying traditional tests for unconditional stationary equilibria, or cointegrating equilibria, in conditions where they are certain about the features of their data. Analysts should use standard critical values in cases where they *know* all their series are stationary. However, if y is $I(1)$, the standard critical values are not the right critical values. Finally, the power comparison between the bounds approach and standard critical values becomes moot when analysts cannot verify their characterizations of their series. When analysts apply standard critical values, they are making strong claims about their knowledge of the series, and these should be explicit.

As Webb, Linn, and Lebo (2019) note, there is an additional advantage of focusing on the LRM when we have multiple independent variables in the model. If y is a unit root, rejecting the null $H_0 : \alpha_1^* = 0$ implies y has an LRR with at least one element of X, but it does not tell us which particular independent variables matter and which do not. But, unlike α_1^*, the LRM is estimated separately for each variable in the model. Thus, the LRM test allows us to draw inferences separately about whether there is an LRR between y and *each* element of X.

The LRM bounds approach is a transparent approach to testing for LRRs because it incorporates the uncertainty in pretesting for all of the series into the hypothesis-testing procedure. It represents a solution to the fundamental problem with standard approaches to applied time-series analysis and allows analysts to move beyond the unit root question in cases, common in applied work, where series cannot be easily classified as $I(0)$ or $I(1)$.

Examples: The LRM Bounds in Action

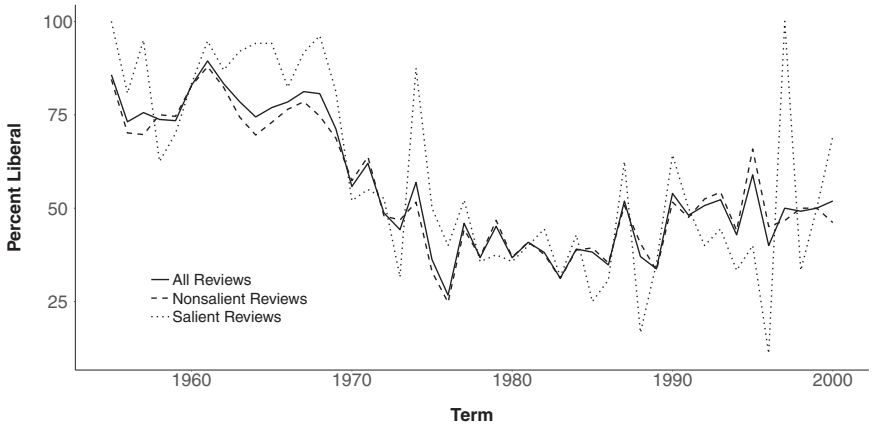
We present two applications of the LRM bounds procedure that demonstrate how it is applied and highlight its empirical relevance. In each case, the answer to the unit root question is ambiguous. This illustrates the need for the bounds procedure. These examples show how testing for LRRs using the LRM bounds testing framework reduces our confidence in the existence of some LRRs established in the literature while enhancing our confidence about others.

Revisiting Casillas, Enns, and Wohlfarth (2011)

How responsive is the Supreme Court to public opinion? This is an age-old and important question (Segal and Spaeth 2002; Stimson, MacKuen, and Erikson 1994). In their analysis, Casillas, Enns, and Wohlfarth (CEW; 2011) use yearly data on Supreme Court decisions from 1956 to 2000 to investigate the short- and long-term effects of public opinion on the ideological direction of the Court's decisions. Using a GECM, the authors estimate the percent of liberal Court reversals as a function of *public policy mood*, *court ideology*, and *social forces*. The authors measure liberal Court reversals separately for *all reviews*, *nonsalient reviews*, and *salient reviews*. Using standard inferential tools applied to the LRM, CEW find evidence of statistically significant LRRs between public policy mood and the percent of liberal reversals among all reviews and nonsalient reviews. The authors conclude that "the public mood directly constrains the justices' behaviors and the Court's policy outcomes, even after controlling for the social forces that influence the public and the Supreme Court," and "the public's awareness of the Court's behavior and the justices' incentives to consider the context of public opinion may be greater than previously thought" (2011, 86).

The authors choose not to report unit root or stationarity test results, noting that the GECM may be estimated in either case. As we have shown above, the appropriate tools for inference depend on the univariate properties of the data. Below, we examine the evidence for and against the presence of a unit root in liberal reversals and demonstrate the ambiguity in the tests. We replicate CEW's analysis and apply the LRM t-statistic

FIGURE 3 Percentage of Liberal Decisions in U.S. Supreme Court Reversals by Type of Case, 1955–2000



Note: The Supreme Court reversals data are taken from Casillas, Enns, and Wohlfarth (2011). *All Reviews* is the percentage of liberal decisions each term among all the cases that reversed lower court rulings. *Salient reviews* is the percentage of liberal decisions each term among salient cases that reversed lower court rulings. *Nonsalient reviews* is the percentage of liberal decisions each term among all nonsalient cases that reversed lower court rulings.

critical value bounds to draw inferences about the existence of an LRR between public policy mood, as well as the other variables in the model, and the percent of liberal reversals.

No theoretical arguments, that we are aware of, have been made regarding the question of whether the ideological balance of Court reversals over time is a unit root process. It is, however, reasonable to think that the ideological balance of decisions is inertial, as a small and relatively constant set of actors is making decisions in each term. In contrast, there is no reason to expect any sort of deterministic trend in the process. As Figure 3 shows, the series show no tendency to grow over time.[17]

Table 3 presents a battery of unit root tests for the percentage of liberal reversals for each type of review. Based on our theoretical understanding of the processes and the evidence in Figure 3, we set $D_t = (1, 0)$ and thus limit consideration to (1) tests of the null hypothesis of a unit root (with neither trend nor drift) against the alternative that the process is mean stationary and (2) the null hypothesis that the series is mean stationary.[18] For both all reviews and nonsalient reviews, the evidence is generally consistent; we cannot reject any of the unit root tests at $\alpha = .05$. However, given the low power of these tests and the very small sample period, the tests are predisposed to fail to reject, regardless of the accuracy of that conclusion. The results of the KPSS test of the mean stationary null are ambiguous: Using short lag truncation, the test rejects the null, but using longer lag truncation inference depends on whether $\alpha = .05$ or $\alpha = .10$ is adopted. The evidence for a unit root in salient reviews is harder to evaluate. Whereas the Dickey-Fuller and DF-GLS tests fail to reject the unit root null, evidence from all versions of the Phillips-Perron test contradict this inference. The results from the KPSS test are identical to those for the other two time series. We conclude that there is enough uncertainty about the nature of the dynamics of these processes that it would be dangerous to use standard normal critical values to test hypotheses on the LRM.

CEW estimated a GECM of the percentage of liberal reversals in which social forces were used to instrument both levels of and changes in Martin-Quinn ideology scores. We have replicated this instrumental variables analysis in Table 4.[19] Standard statistical inference based

[17]The first and last observations have disproportionate influences in diagnostic regressions. It is thus possible one could infer a trend exists, but these results would likely change if the sampling window changed.

[18]This decision does not affect inferences. See the supporting information for further details (Section 1, 4–8).

[19]Analysis is conducted using the `dynlm` package in `R`. Estimates are identical to those reported by CEW, who used Stata. Calculation of the variance–covariance matrix is different in the two packages such that our standard errors are larger than those reported by CEW. In addition, CEW do not explain how they generated standard errors

TABLE 3 Unit Root and Stationarity Tests: Liberal Decisions (Casillas, Enns, and Wohlfarth 2011), 1955–2000

	Variable		
Test	**All Reviews**	**Nonsalient Reviews**	**Salient Reviews**
Dickey-Fuller[a] τ_μ	-2.77^{+}	-2.44	-1.75
Phillips-Perron Z_τ (short)	-2.10	-2.15	-3.99^{**}
Phillips-Perron Z_τ (long)	-2.20	-2.24	-4.65^{**}
Phillips-Perron Z_α (short)	-6.16	-6.64	-23.07^{**}
Phillips-Perron Z_α (long)	-7.14	-7.50	-35.50^{**}
DF-GLS[b]	-0.82	-1.12	-0.82
KPSS μ (short)	0.83^{**}	0.79^{**}	0.86^{**}
KPSS μ (long)	0.39^{+}	0.38^{+}	0.42^{+}

Note: $T = 46$. The null hypothesis for the Dickey-Fuller, Phillips-Perron, and DF-GLS tests is that the series is a unit root; the alternative hypothesis is that the series is mean stationary. Short lag truncation is based on the formula $trunc(4 * (n/100)^{0.25}) = 3$; long lag truncation is based on the formula $trunc(12 * (n/100)^{0.25}) = 9$. Further details are given in the supporting information (Section 1, 4–6).
[a]Results are from a model with the lag length selected using a general-to-specific modeling strategy and $\alpha = .05$, beginning with a maximum lag length of 9. Eight lags were selected for both *All Reviews* and *Nonsalient Reviews*. Two lags were selected for *Salient Reviews*. The AIC selected two lags in all cases. Inferences do not depend on this decision.
[b]Results are based on a lag length selected using a general-to-specific modeling strategy: one lag for *All Reviews*, eight lags for *Nonsalient Reviews* and *Salient Reviews*.
$^{**}p < .01$, $^{*}p < .05$, $^{+}p < .10$.

on these results suggests, as CEW report, policy mood, Court ideology, and social forces have large short-term effects on the percentage of liberal revisions, although the estimated effect of policy mood is less than its standard error in the model of salient reviews.[20] The effect of lagged levels of these variables is a considerably mixed bag. In particular, policy mood does not reach standard levels of significance in any equation.

If we are interested in whether these variables drive liberal reversals in the long run, our focus should be on the LRMs. And given our uncertainty about the univariate dynamics of the dependent variables, we should be wary of inferences based on standard normal distributions. Thus, we apply the LRM critical bounds to evaluate the CEW hypotheses. The bottom portion of Table 4 presents two sets of information for each model and variable. In the first, third, and fifth columns are the LRMs $(-\beta_1^*/\alpha_1^*)$ and their standard errors. The standard errors were calculated using the delta method. The second, fourth, and sixth columns report the *t*-statistics and inferences based upon the $\alpha = .05$ bounds. Given $T = 45$ and three independent variables, the most applicable bounds are 1.10 and 3.64. In only one case is the LRM *t*-statistic above the upper bound such that we can reject the null of no LRR: The baseline Court ideology score is tied significantly to the disposition of the full set of reviews. Three of the LRM *t*-statistics are below the bounds: policy mood in the model of salient reviews and social forces in the full set of reviews and nonsalient reviews. We fail to reject the null of no LRR in these cases; mood does not constrain the Supreme Court on salient cases. This is consistent with CEW's conclusions. The remaining LRM *t*-statistics are well inside the bounds such that we are uncertain whether there is an LRR absent certainty about the univariate dynamics. Thus, we caution against inferring that public opinion influences the Court's decisions, even in nonsalient cases.

for the LRM, and we were unable to replicate them. Given that CEW's GECMs are IV regressions, we used the delta method to calculate the standard errors for the LRMs. These are larger than those reported by CEW.

[20]If the dependent variables contain unit roots, inferences on these variables are nonstandard. Sims, Stock, and Watson (1990) have shown that if y and any element of X are integrated, standard limiting distributions only apply to hypothesis tests on coefficients that can be written as coefficients on mean zero stationary variables and only if the model is correctly specified such that the errors are uncorrelated. In our examples, hypothesis tests on β_0 and β_1 in the ADL and β_0^* in the GECM can be evaluated using standard limiting distributions given that there is no evidence the errors are correlated. However, standard limiting distributions do not apply to the *F*-test on $\beta_0 + \beta_1$ in the ADL or *t*-test on β_1^* in the GECM, unless the data are cointegrated.

Explaining Labour Party Support in the UK, 1997–2010

Our second example looks at monthly vote intentions for the Labour Party in the United Kingdom for the period May 1997 to April 2010, as a function of prime ministerial approval, economic optimism, and approval of the opposition leader. This example is different in four ways.

TABLE 4 Casillas, Enns, and Wohlfarth's (2011) GECMs of Supreme Court Reversals, 1956–2000

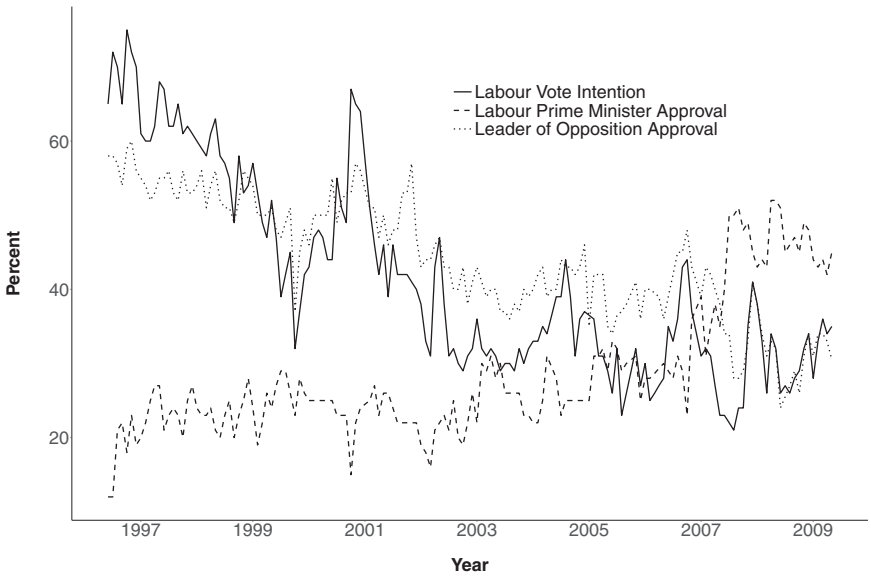
Variables	All Reviews	Nonsalient Reviews	Salient Reviews
$\alpha_1^* y_{t-1}$			
Lagged Reviews	−0.83	−0.77	−1.27
	(0.16)	(0.16)	(0.17)
$\beta_0^* \Delta x_t$			
Public Policy Mood	1.59	1.68	1.24
	(0.86)	(0.87)	(1.99)
Court Ideology	12.48	11.37	10.47
	(4.73)	(4.87)	(11.07)
Social Forces(IV)	2.78	2.56	7.49
	(3.31)	(3.29)	(8.86)
$\beta_1^* x_{t-1}$			
Public Policy Mood	0.87	0.88	0.71
	(0.46)	(0.47)	(1.06)
Court Ideology	9.68	8.37	16.90
	(3.42)	(3.41)	(6.32)
Social Forces	0.98	−0.01	9.21
	(2.37)	(2.37)	(5.61)
Constant	−6.03	−11.05	31.02
	(27.73)	(28.04)	(64.32)
Long-Run Multiplier			
	LRM (s.e.) / **LRM *t* (Inference)**	**LRM (s.e.)** / **LRM *t* (Inference)**	**LRM (s.e.)** / **LRM *t* (Inference)**
Public Policy Mood	1.05 (0.56) / 1.86 (Between)	1.15 (0.62) / 1.85 (Between)	0.56 (0.84) / 0.67 (Below)
Court Ideology	11.67 (3.11) / 3.75 (Beyond)	10.90 (3.43) / 3.18 (Between)	13.36 (4.65) / 2.87 (Between)
Social Forces	1.18 (2.81) / 0.42 (Below)	−0.01 (3.08) / 0.00 (Below)	7.28 (4.23) / 1.72 (Between)
N	45	45	45
R^2	0.53	0.49	0.64
Breusch-Godfrey (4 lags)	0.433	0.361	0.487

Note: Standard errors are in parentheses. The standard errors of the LRMs were calculated using the delta method. The *t*-statistics are reported as "Below" when $|t| < 1.10$, "Between" when $1.10 < |t| < 3.64$, and "Beyond" when $|t| > 3.64$.

First, the series are longer ($T = 155$ months). Second, the dependent variable in this example has a gradual but distinct downward tendency. This could be caused by a constant (drift) in the DGP or by a trend in the DGP. Third, to show the flexibility of the LRM bounds test, we begin by estimating an autoregressive distributed lag (ADL) model rather than a GECM and go through the different calculations one should use in that approach. And, fourth, based on previous research, we expect to find an LRR between at least one of the independent variables—party leader support—and the dependent variable.

Figure 4 shows the dependent variable, *Labour Vote Intentions*, and two of the three independent variables of interest, *Prime Minister Approval* and *Leader of Opposition Approval*. Under the leadership of Tony Blair, the Labour Party swept to victory in May 1997, but the ratings of both the party and the prime minister fell precipitously in the years that followed. In particular, the fallout over Britain's involvement in the Iraq War pushed Blair's approval rating into the 20s and led to his replacement by Gordon Brown in 2007. Prior studies have shown a close relationship between party vote intentions and leadership ratings. Clarke, Stewart, and Whiteley (1997) find

FIGURE 4 Vote Intentions for Labour, Prime Ministerial Approval, and Leader of Opposition Approval, May 1997–April 2010



Note: Data come from Market and Opinion Research International's (MORI) social trends (https://www.ipsos.com/ipsos-mori/en-uk/social-trends). *Labour vote intention* is the monthly percentage of vote intention for the Labour Party. *Labour Prime Minister Approval* is the percent satisfied with the way the prime minister is running the country. *Leader of Opposition Approval* is the percent satisfied with the leader of the opposition.

cointegration between the prime minister's ratings and his or her party's vote intentions, and Lebo and Young (2009) find fractional cointegration between party and leadership ratings in both the governing party and the official opposition.

The univariate diagnostics are complicated by the downward trajectory in the dependent variable.[21] Results from unit root tests depend on the analyst's decision whether to entertain a deterministic trend. Including a trend allows rejection of the null hypothesis of a unit root with drift in favor of the alternative that the series is trend stationary. Without including a trend, however, we fail to reject the unit root null. For the independent variables, with or without a trend, the conclusions are ambiguous and depend on one's choice of test.[22] Given these results, traditional approaches to time-series analysis will be problematic. In particular, results will be suspect if one tries to use these tests to decide whether the series are stationary, which model to estimate, and, finally, how one should interpret α_1^*.

We begin by estimating an ADL model with a trend. Long-run multipliers are calculated for each independent variable ($\frac{(\beta_0+\beta_1)}{(\alpha_1-1)}$), and the standard errors are equivalently estimated using the delta method and the Bewley IV regression. Table 5 shows the results. The results from the ADL can be interpreted at face value and show a significant effect for prime ministerial approval (0.392 with s.e. = 0.051).

Looking at Table 2, the bounds for T = 150 and k = 3 are 1.01 and 3.65 at the .05 level. The LRM test statistic for *economic optimism* is between the bounds, giving us an inconclusive result. To draw a firm conclusion about economic optimism, we would need to know more about the dynamic properties of the data. On the other hand, the LRM test statistics for approval of both the prime minister and leader of the opposition are each well above both bounds such that we can reject the null hypotheses of no LRRs with Labour vote intentions with confidence.

[21]This example also illustrates that the univariate nature of a series can depend on the time period chosen. Over a period that includes multiple governments, we see popularity rise and fall with no trend evident across the full range of data. However, during a period of one-party government, the inexorable descent described by Mueller (1970) is evident, even if it is not a *deterministic* trend.

[22]Extended unit root test results can be found in the supporting information (Section 1, 9–12).

TABLE 5 ADL Model of Labour Vote Intention, May 1997–April 2010

Labour Vote Intentions	ADL	LRM x_{it}	LRM$-t$
$\alpha_1 y_{t-1}$			
Labour Vote Intention	0.392		
	(0.076)		
$\beta_0 x_t$			
Prime Minister Approval	0.309	0.396	7.68
	(0.051)	(0.052)	(Beyond)
Economic Optimism	−0.009	−0.050	−2.35
	(0.025)	(0.021)	(Between)
Leader of Opposition Approval	−0.077	−0.329	−4.93
	(0.067)	(0.067)	(Beyond)
$\beta_1 x_{t-1}$			
Prime Minister Approval	−0.069		
	(0.058)		
Economic Optimism	−0.021		
	(0.025)		
Leader of Opposition Approval	−0.123		
	(0.067)		
Trend	−0.014		
	(0.012)		
Constant	22.986		
	(3.489)		
N	155		
R^2	0.912		
Breusch-Godfrey (12 lags)	0.831		

Note: Standard errors are in parentheses. The LRM, LRM_{SE}, and t-LRM, are estimated from Equation (4), the Bewley instrumental variables regression. The t-statistics are reported as "Below" when $|t| < 1.01$, "Between" when $1.01 < |t| < 3.65$, and "Beyond" when $|t| > 3.65$.

Crucially, this confidence does not rely on particular conclusions about the univariate properties of the variables. This highlights the key advantage of the LRM bounds procedure over standard approaches.

Beyond the Unit Root Question

Time-series texts present a seemingly simple blueprint for applied time-series analysis. Standard approaches begin with a set of pretests that are supposed to neatly classify series as either unit root or stationary processes. These diagnoses, then, determine the strategies one should use to test for LRRs. These approaches are built on weak foundations. The pretests are notoriously unreliable and often produce conflicting results. In the face of uncertainty, the ambivalent analyst is forced to make definitive choices. This uncertainty is not reflected in final analyses. The standard blueprint does not allow the analyst to be uncertain about the univariate properties of his or her data. If the foundation for an analysis is weak, the final result may be little more than a house of cards. We need a new blueprint.

The LRM bounds procedure developed by Webb, Linn, and Lebo (2019) and described in this article provides a principled approach to inference that explicitly incorporates analysts' uncertainty about the univariate properties of their data into hypothesis testing. If analysts misclassify their series, they will draw incorrect inferences about LRRs. This is the fundamental problem with traditional approaches. The LRM bounds procedure provides an elegant solution. The LRM t-test is valid for both stationary and cointegrating equilibria, and the critical value bounds encompass all possible critical values that could be correct given any possible combination of $I(0)$ or $I(1)$ variables. This simplifies time-series analysis by obviating the need for pretesting and allows analysts to move beyond the unit root question.

For a discipline whose primary interest in time-series analysis is hypothesis testing, the LRM bounds procedure represents a critical innovation. In closing, we offer a new blueprint.

- **Theory**: Time-series analysis still begins with theory. Theory determines which variables should be included in the model. Analysts must choose models that are general enough to encompass their theories. Analysts derive hypotheses from their theories that speak to the existence of LRRs.
- **Estimation and Inference**: Analysts should test their hypotheses about LLRs by comparing the LRM t-statistics to the appropriate bounds. This takes uncertainty about the univariate dynamics into account and allows analysts to reach reliable conclusions about the existence of LLRs regardless of univariate characteristics. In many cases, this is the primary goal of applied time-series analyses.
- **Classification and Interpretation**: Analysts may want to classify their equilibrium relationships. This can be extremely difficult, especially with short time series. This should be done cautiously and requires that analysts have long time series, consistent results from unit root and stationarity tests, and strong theory. Unit root tests do not need to be thrown out entirely, but they should be applied with the appropriate skepticism.
- **Reporting Results**: Analysts should strive for transparency when reporting results. If analysts classify equilibrium relationships, they should report a broad range of results, including (1) results from multiple tests, (2) results that support their specification of D_t, and (3) the level of significance associated with each test. If tests produce competing results; theory, our knowledge of the data, and our knowledge of the testing procedures should be used to make a case for one conclusion over another.

The advice of 40 years of time-series practice is to begin by doing our best to diagnose the properties of data but then to ignore the uncertainty in those decisions in the models and inferences that follow. Standard approaches require analysts to be certain over what is inherently uncertain. The LRM bounds testing procedure provides a unified way of thinking about dynamic specification that avoids the pitfalls common to alternative procedures. The practices we outline here are far less likely to produce misleading conclusions and also make it less likely that important substantive debates will stall over disagreements about basic time-series principles. As time-series practitioners move beyond the unit root question, they can concentrate on pushing important theoretical developments forward.

References

Banerjee, Anindya, Juan Dolado, John W. Galbraith, and David F. Hendry. 1993. *Co-integration, Error Correction, and the Econometric Analysis of Non-Stationary Data.* Oxford: Oxford University Press.

Bewley, R. A. 1979. "The Direct Estimation of the Equilibrium Response in a Linear Model." *Economic Letters* 3(4): 357–61.

Box-Steffensmeier, Janet M., John R. Freeman, Matthew P. Hitt, and Jon C. W. Pevehouse. 2014. *Time Series Analysis for the Social Sciences.* New York: Cambridge University Press.

Box-Steffensmeier, Janet M., and Andrew R. Tomlinson. 2000. "Fractional Integration Methods in Political Science." *Electoral Studies* 19(1): 63–76.

Brandt, Patrick T., and John R. Freeman. 2009. "Modeling Macro-Political Dynamics." *Political Analysis* 17(1): 113–42.

Burke, Simon, and John Hunter. 2005. *Modelling Non-Stationary Economic Time Series: A Multi-variate Approach.* New York: Springer.

Campbell, John Y., and Pierre Perron. 1991. "Pitfalls and Opportunities: What Macroeconomists Should Know about Unit Roots." *NBER Macroeconomics Annual* 6: 141–201.

Casillas, Christopher J., Peter K. Enns, and Patrick C. Wohlfarth. 2011. "How Public Opinion Constrains the U.S. Supreme Court." *American Journal of Political Science* 51(1): 74–88.

Choi, In. 2015. *Almost All about Unit Roots: Foundations, Developments, and Applications.* New York: Cambridge University Press.

Clarke, Harold D., and Matthew Lebo. 2003. "Fractional (Co)Integration and Governing Party Support in Britain." *British Journal of Political Science* 33(2): 283–301.

Clarke, Harold D., and Marianne C. Stewart. 1994. "Prospections, Retrospections and Rationality: The 'Bankers' Model of Presidential Approval Reconsidered." *American Journal of Political Science* 38(4): 1104–23.

Clarke, Harold D., Marianne C. Stewart, and Paul Whiteley. 1997. "Tory Trends: Party Identification and the Dynamics of Conservative Support Since 1992." *British Journal of Political Science* 27(2): 299–331.

De Boef, Suzanna, and Jim Granato. 1997. "Near-Integrated Data and the Analysis of Political Relationship." *American Journal of Political Science* 41(2): 619–40.

De Boef, Suzanna, and Luke Keele. 2008. "Taking Time Seriously." *American Journal of Political Science* 52(1): 184–200.

De Boef, Suzanna, and Paul Kellstedt. 2004. "The Political (and Economic) Origins of Consumer Confidence." *American Journal of Political Science* 38(4): 633–49.

DeJong, David N., John C. Nankervis, N. Eugene Savin, and Charles H. Whiteman. 1992. "The Power Problems of Unit Root Tests in Time Series with Autoregressive Errors." *Journal of Econometrics* 53(1–3): 323–43.

Dickey, David A., and Wayne A. Fuller. 1979. "Distribution of the Estimators for Autoregressive Time Series with a Unit Root." *Journal of the American Statistical Association* 74(366a): 427–31.

Elliott, Graham, Thomas J. Rothenberg, and James H. Stock. 1996. "Efficient Tests for an Autoregressive Unit Root." *Econometrics* 64(4): 813–836.

Enders, Walter. 2015. *Applied Econometric Time Series.* 4th ed. New York: Wiley and Sons.

Engle, Robert F., and C. W. J. Granger. 1987. "Co-integration and Error Correction: Representation, Estimation, and Testing." *Econometrica* 55(2): 251–76.

Enns, Peter K., Nathan J. Kelly, Takaaki Masaki, and Patrick C. Wohlfarth. 2016. "Don't Jettison the General Error Correction Model Just Yet: A Practical Guide to Avoiding Spurious Regression with the GECM." *Research & Politics* 3(2): 1–16.

Enns, Peter K., Nathan J. Kelly, Takaaki Masaki, and Patrick C. Wohlfarth. 2017. "Moving Forward with Time Series Analysis." *Research & Politics* 4(4): 1–7.

Enns, Peter K., and Christopher Wlezien. 2017. "Understanding Equation Balance in Time Series Regression." *The Political Methodologist* 24(2): 2–12.

Ericsson, Neil R., and James G. MacKinnon. 2002. "Distributions of Error Correction Tests for Cointegration." *Econometrics Journal* 5(2): 285–318.

Esarey, Justin. 2016. "Fractionally Integrated Data and the Autoregressive Distributed Lag Model: Results from a Simulation Study." *Political Analysis* 24(1): 42–49.

Evans, G. B. A., and N. Eugene Savin. 1981. "Testing for Unit Roots: 1." *Econometrica: Journal of the Econometric Society* 49(3): 753–79.

Evans, G. B. A., and N. Eugene Savin. 1984. "Testing for Unit Roots: 2." *Econometrica: Journal of the Econometric Society* 52(5): 1241–69.

Grant, Taylor, and Matthew J. Lebo. 2016. "Error Correction Methods with Political Time Series." *Political Analysis* 24(1): 3–30.

Helgason, Agnar Freyr. 2016. "Fractional Integration Methods and Short Time Series: Evidence from a Simulation Study." *Political Analysis* 24(1): 59–68.

Hendry, David F. 1995. *Dynamic Econometrics.* Oxford: Oxford University Press.

Juhl, Ted, and Zhijie Xiao. 2003. "Power Functions and Envelopes for Unit Root Tests." *Econometric Theory* 19(2): 240–53.

Keele, Luke, Suzanna Linn, and Clayton McLaughlin Webb. 2016a. "Concluding Comments." *Political Analysis* 24(1): 83–86.

Keele, Luke, Suzanna Linn, and Clayton McLaughlin Webb. 2016b. "Treating Time with All Due Seriousness." *Political Analysis* 24(1): 31–41.

Kwiatkowski, Denis, Peter C. B. Phillips, Peter Schmidt, and Yongcheol Shin. 1992. "Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root." *Journal of Econometrics* 54(1–3): 159–78.

Lebo, Matthew J., and Taylor Grant. 2016. "Equation Balance and Dynamic Political Modeling." *Political Analysis* 24(1): 69–82.

Lebo, Matthew J., and Patrick W. Kraft. 2017. "The General Error Correction Model in Practice." *Research & Politics* 4(2): 1–13.

Lebo, Matthew J., and Andrew J. O'Geen. 2011. "The President's Role in the Partisan Congressional Arena." *Journal of Politics* 73(3): 718–34.

Lebo, Matthew J., and Everett Young. 2009. "The Comparative Dynamics of Party Support in Great Britain: Conservatives, Labour and the Liberal Democrats." *Journal of Elections, Public Opinion and Parties* 19(1): 73–103.

Mueller, John. 1970. "Presidential Popularity from Truman to Johnson." *American Political Science Review* 65(1): 18–34.

Müller, Ulrich K., and Graham Elliott. 2003. "Tests for Unit Roots and the Initial Condition." *Econometrica* 71(4): 1269–86.

Ng, Serena, and Pierre Perron. 1995. "Unit Root Tests in ARMA Models with Data–Dependent Methods for the Selection of the Truncation Lag." *Journal of the American Statistical Association* 90(429): 268–81.

Ng, Serena, and Pierre Perron. 2001. "Lag Length Selection and the Construction of Unit Root Tests with Good Size and Power." *Econometrica* 69(6): 1519–54.

Pagan, Adrian. 1987. "Three Econometric Methodologies: A Critical Appraisal 1." *Journal of Economic Surveys* 1(1–2): 3–23.

Perron, Pierre. 1989. "The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis." *Econometrica: Journal of the Econometric Society* 57(6): 1361–1401.

Perron, Pierre, and Serena Ng. 1996. "Useful Modifications to Some Unit Root Tests with Dependent Errors and Their Local Asymptotic Properties." *Review of Economic Studies* 63(3): 435–63.

Pesaran, M. Hashem, Yongcheol Shin, and Richard J. Smith. 2001. "Bounds Testing Approaches to the Analysis of Level Relationships." *Journal of Applied Econometrics* 16(3): 289–326.

Philips, Andrew Q. 2018. "Have Your Cake and Eat It Too? Cointegration and Dynamic Inference from Autoregressive Distributed Lag Models." *American Journal of Political Science* 62(1): 230–44.

Phillips, Peter C. B., and Pierre Perron. 1988. "Testing for a Unit Root in Time Series Regression." *Biometrika* 75(2): 335–46.

Pickup, Mark, and Paul Kellstedt. 2020. "Equation Balance in Time Series Analysis: What It Is and How to Apply It." Working Paper. (January 28, 2020). Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3526534.

Podivinsky, Jan M., and Maxwell L. King. 2000. "The Exact Power Envelope of Tests for a Unit Root." Working Paper. https://eprints.soton.ac.uk/32895/.

Schwert, G. William. 1989. "Tests for Unit Roots: A Monte Carlo Investigation." *Journal of Business & Economic Statistics* 7(2): 147–59.

Segal, Jeffrey A., and Harold J. Spaeth. 2002. *The Supreme Court and the Attitudinal Model Revisited.* New York: Cambridge University Press.

Sims, Christopher A., James H. Stock, and Mark W. Watson. 1990. "Inference in Linear Time Series Models with Some Unit Roots." *Econometrica: Journal of the Econometric Society* 58(1): 113–44.

Stimson, James A., Michael B. MacKuen, and Robert S. Erikson. 1994. "Opinion and Policy: A Global View." *PS: Political Science and Politics* 27(1): 29–35.

Stock, James H. 1991. "Confidence Intervals for the Largest Autoregressive Root in U.S. Macroeconomic Time Series." *Journal of Monetary Economics* 28(3): 435–59.

Webb, Clayton McLaughlin, Suzanna Linn, and Matthew Lebo. 2019. "A Bounds Approach to Inference Using the Long Run Multiplier." *Political Analysis* 27(3): 1–21.

Williams, John T. 1992. "What Goes Around Comes Around: Unit Root Tests and Cointegration." *Political Analysis* 4: 229–35.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information